Multilabel Tagging of Discourse Relations in Ambiguous Temporal Connectives

Yannick Versley
Collaborative Research Centre (SFB) 833
University of Tübingen
versley@sfs.uni-tuebingen.de

Abstract

Many annotation schemes for discourse relations allow combinations such as *temporal+cause* (for events that are temporally and causally related to each other) and *temporal+contrast* (for contrasts between subsequent time spans, or between events that are temporally coextensive). However, current approaches for the automatic classification of discourse relations are limited to producing only one relation and disregard the others.

We argue that the information contained in these 'additional' relations is indeed useful and present an approach to tag multiple fine-grained discourse relations in ambiguous connectives from the German TüBa-D/Z corpus. Using a rich feature set, we show that good accuracy is possible even for inferred relations that are not part of the connective's 'core' meaning.

1 Introduction

In order to account for the structure of text beyond the level of single clauses, it is common to postulate *discourse relations* holding between clauses or groups of clauses. Discourse relations are frequently marked by *connectives* such as *because*, *as* or *while*, which give an indication both of (syntactic or anaphoric) linking possibilities for the spans and of the possible relations.

Many connectives (such as *because* or *for instance*) always signal one specific discourse relation. This fact has, after initial successes in purely structural discourse parsing (Soricut and Marcu, 2003), led to decreased attention from researchers. Other connectives, however, are ambiguous between multiple readings and their disambiguation necessitates similar semantic information as implicit (connective-less) discourse relations.

Ambiguous temporal markers such as *after*, *as* or *while* usually occur with a purely temporal reading, but also with additional non-temporal discourse relations, such as causal and contrastive readings. When these non-temporal relations occur instead of, or in addition to, the temporal reading, they require similar similar inferences from the reader as in connective-less discourse relations, but may be easier to detect automatically. For our goal of accurate classification, multilabel classification becomes necessary when the non-temporal discourse relations co-occur with the temporal ones:

(1) a. As [$_{\text{arg2}}$ *individual investors have turned away from the stock market over the years*], [$_{\text{arg1}}$ *securities firms have scrambled to find new products that brokers find easy to sell*].

b. [$_{\text{arg1}}$ *"Forget it," he said*] as [$_{\text{arg2}}$*he handed her a paper*].

c. But as [$_{\text{arg2}}$ *the French embody a Zen-like state of blase when it comes to athletics*] (try finding a Nautilus machine in Paris), [$_{\text{arg1}}$ *my fellow conventioners were having none of it*].

In the examples from (1), the sentence in (b) is clearly temporal (and non-causal), and the one in (c) is clearly causal (and non-temporal), whereas in (a) the connective contributes both a causal and a temporal aspect to the coherence of the text.

In the Penn Discourse Treebank (Prasad et al., 2008), which uses multiple labels as a last resort

Proceedings of Recent Advances in Natural Language Processing, pages 154–161,
Hissar, Bulgaria, 12-14 September 2011.

when annotators cannot reach an agreement or feel that an instance is inherently ambiguous, 5.5% of discourse connectives are assigned multiple discourse relations. The proportion of multiple vs. single discourse relation varies from connective to connective, with a higher proportion in ambiguous temporal connectives, where it ranges from *after*'s 9% and *while*'s 12.7% over *as* (23.6%) and *when* (21%) to *meanwhile* with 70% of the instances that have multiple labels.

The annotation of discourse connectives in the TüBa-D/Z (Telljohann et al., 2009), which we used in our experiments, uses combinations of temporal and other relations to signal causation between successive events or a contrast between co-temporal events, yielding 64.6% of multilabel instances for *nachdem* (after/since), and 53.8% of multilabel instances for *während* (while).

Hence, it is necessary for accurate classification to identify *both* of the discourse relations holding in such a case, whereas most recent research, such as Pitler and Nenkova (2009) or Wellner (2009) has focused on single-relation classification.[1]

A notable exception is Bethard and Martin's (2008) work on instances of *and*, where the presence of a temporal or causal relation is classified independently of the other.

In terms of the features used in classification, the perception that most connectives are unambiguous has created a disparity in terms of features between approaches that target discourse relations signaled by a connective (so-called *explicit* relations) and those that are inferred between adjacent discourse segments in the absence of connectives (*implicit* relations).

Work on explicit (i.e., connective-bearing) relations has emphasized simpler features, such as the syntactic neighbourhood of the connective (Pitler and Nenkova, 2009) or features based on tense and mood of the argument clauses (Miltsakaki et al., 2005). In contrast, work targeting implicit discourse relations harnesses a larger variety of features, including word pairs (Marcu and Echihabi, 2002; Sporleder and Lascarides, 2008), structural properties of the argument clauses (Lin et al., 2009), semantic parallelism between arguments' main verbs' classes, emotive polarity, and other special word categories (Pitler et al., 2009).

In the remainder of this paper, we formulate the disambiguation of ambiguous temporal connectives as a multilabel classification task (where the system can, and should, assign more than one discourse relation). The results (sections 5, 6) show that a rich feature set - partly inspired by the state of the art for implicit relations - is instrumental in detecting the 'non-obvious' discourse relations in temporal connectives.

2 Annotating Ambiguous Temporal Connectives in the TüBa-D/Z

For our study on automatic classification, we use instances of two German temporal connectives that can also carry a non-temporal discourse relation, namely *während* and *nachdem*:

The default reading of *nachdem* (corresponding to English *after/as/since*) signals a *temporal* relation between subsequent events, which is also compatible with a *causal* discourse relation, or a *contrast* between two events or states. *Nachdem* is also used in contexts where it confers an argumentative relation between propositions (*evidence*), or between a licensing proposition and a question or imperative (*speech-act*). As seen in example (1), *rhetorical* relations such as 'evidence' and 'speech-act' can occur with arguments that would be incompatible with the temporal reading of *nachdem*:

(2) Und *nachdem* ja die vertraglichen Bindungen noch weiterlaufen, und zwar bis zum Jahre 2006, werden heuer und in den kommenden Jahren noch weitere 250 Millionen Euro zur Auszahlung gelangen.
And as *the contractual obligations are still in force, and run up to 2006, this year and in the coming years a further EUR 250 million will be paid out.*

Similar to its English counterpart *while*, German *während* has a *temporal* reading that locates the sub-clause in the phase of the matrix clause, but also allows a *contrast* reading where two propositions are contrasted with respect to a common integrator.

In a prototypical example such as (3), we find a parallel structure with one pair of entities being compared (*Mary* and *Peter*) and an attribute in which they differ (liking *bananas* versus prefering

[1] Both Pitler and Nenkova, and Wellner classify only the first relation. Pitler and Nenkova count the system response as correct when it includes any of the discourse relations in the gold standard, while Wellner counts a system-generated relation as correct if it reproduces the first of the two relations of a multi-relation instance.

Relations	*nachdem*	*während*
Temporal	93.9	76.7
Result	60.2	
├ situational	*53.4*	
│ ├ enable	*31.6*	
│ └ cause	*21.7*	
└ rhetorical	*6.4*	
├ evidence	*4.1*	
└ speech-act	*2.4*	
Comparison	10.5	76.7
├ parallel	*4.8*	
└ contrast	*5.8*	*76.7*

Percent of instances tagged with a given label (including subcategories); Numbers across top-level relations sum up to more than 100% because of multi-label instances.

Table 1: Discourse relation inventory

peaches).

(3) Während [Maria] [Bananen] mag, bevorzugt [Peter] [Pfirsiche].
While [*Mary*] *likes* [*bananas*], [*Peter*] *prefers* [*peaches*].

Such a structure, which we can describe using a common integrator such as "*People* like *fruits*", receives the *contrast* relation.

In cases where a contrast coincides with cotemporal states, or a temporal relation coincides with an inferred contrast, a secondary temporal or contrast relation is annotated to reflect the ambiguity.

Our data set – the connective occurrences from the current extent of the TüBa-D/Z plus additional texts that are scheduled for the inclusion in one of the next releases, totaling about 60 000 senteces – contains 294 instances of *nachdem* and 527 instances of *während*. Where available, we used the syntactic annotation from the treebank; in the remaining cases, we used a syntactic parser (Versley and Rehbein, 2009) to provide syntax trees for the feature extraction. Table 1 shows the full taxonomy of relations for the ambiguous connectives considered in the experiments.

3 Multilabel classification

Reproducing the connective annotation in the TüBa-D/Z presents a hierarchical multi-label classifcation task: more than one label may apply to a given instance, and labels are arranged in taxonomical categories.

As in classical multi-label tagging, the classifier should take into account the suitability of individual classification labels for a given example; however, the context of discourse relation classification shows stronger interdependence of labels (e.g., a non-temporal example is bound to have an evidence or contrast relation).

3.1 Evaluating multilabel classification

As multilabel classification goes beyond assigning exactly one atomic label, scoring whether the proposed label combination is identical to the gold standard (*equal* in the results table) fails to give partial credit to a system response that reproduces some, but not all of the correct discourse relations.

The *dice* evaluation measure accounts for the overlap between the gold standard label combination and the label combination in the system response, calculated as $\frac{2|A \cap B|}{|A|+|B|}$. Both *equal* and *dice* measure can be calculated at each level of the taxonomy, yielding values for $d = 1$ (the topmost level) up to $d = 3$ (the finest taxonomic level).

In addition, the assignment of any particular relation can be evaluated using the standard F-measure and precision/recall.

3.2 Greedy classification

One of the classical approaches to multilabel classification is to decompose the labeling decision into binary decisions for each possible label (one-vs-all reduction) and using confidence values to choose one or several labels among those that are most confidently classified as positive examples.

To yield the finer-grained distinctions from the taxonomy (such as *Comparison.contrast* vs. *Comparison.parallel*), the classifier makes an additional decision on the fine-grained class corresponding to the coarse-grained one, which is again realized through training separate classifiers for each fine-grained relation.

In our experiments, we use SVMperf, an SVM implementation that is able to train classifiers optimized for performance on positive instances (Joachims, 2005). To improve the separability of the data (SVMperf, like the AMIS package used for CRF training, uses linear classifiers), we use feature combinations up to degree 2.

3.3 A CRF-based approach

One disadvantage of the greedy decomposition into a sequence of binary decisions outlined above is that this variant is unable to model dependencies between the labels assigned by the system; similarly, the greedy decomposition is unable to use evidence for or against individual fine-grained relations in the decision regarding the coarse-grained relations.

As an alternative approach, we consider a classifier that directly ranks possible label combinations, considering all (fine-grained) labels at once. The model ranks all label combinations $Y \in \mathcal{Y}$ using a feature function Φ and the learned weight vector w:

$$\overline{Y} = \arg\max_{Y \in \mathcal{Y}} \langle w, \Phi(x, Y) \rangle$$

where $\mathcal{Y}$ contains all allowable label combinations and Φ extracts a *feature vector* containing the information about the problem instance (x) and the label combination under consideration (Y).

In order to describe each instance, we factor Φ as $\Phi(x, Y) := \Phi_{\text{lab}}(Y) \times \Phi_{\text{data}}(x)$ (i.e., assuming a label feature `Temporal` and a data feature `main-present`, Φ would contain the combined feature (`Temporal,main-present`)).

In our case, the label information from Φ_{lab} contains the set of coarse-grained relations assigned (e.g. `Temporal+Result`), as well as the fine-grained relations, individually (in the example, both `Temporal` and `Result.situational.enable`). It is easy to see that the problem size increases superlinearly with the number of possible relations, because the set $\mathcal{Y}$ of possible labelings can grow quadratically. Keeping the problem size in check provides a gain in efficiency that is already helpful at the current data size, and becomes crucial as the label set and amount of data grow with the addition of more connectives.

To mitigate this problem, we factor the actual feature vector into a *feature forest* (Miyao and Tsujii, 2002) that contains shared nodes for each element, which means that the necessary computations become linear in (*number of fine-grained relations+ number of coarse-grained relation combinations*).

Since the CRF approach optimizes for likelihood of the correct (fine-grained) solution, the results of the CRF classifier may not always give optimal results with respect to a given evaluation metric. To compensate for this, we introduce a *bias* parameter that is added to the score of candidate labelings with more than one label, which forces the classifier towards including (more) labels even when it is not completely certain about them.

4 Classification features

In contrast to newer work in this area, earlier approaches for explicit discourse relations, such as Miltsakaki et al. (2005), have mainly relied on linguistic features indicating the clause or event type, which allows to separate temporal from atemporal uses of a connective in some cases. For our classification experiments, we include a set of baseline features reflecting these linguistic properties as well as more specific features aiming at the differences between different types of argument clauses, but also features that target broader lexical information – in this case, those aimed at the semantics of each argument clause (by taking the head itself, or a characterization), but also co-taxonomic relations between the argument clauses as well as pairs of lemmas and (syntactic) productions.

A first set of *baseline features* include basic linguistic features, such as **clause order** (i.e., topicalization/fronting), as the non-temporal discourse relations are more likely to occur with fronted subclauses than with postposed ones; **tense** features include indicators for perfect, passives, and modal verbs as well as the tense of the finite verb in each clause; a binary **negation** feature indicates the presence of negating adverb (e.g., English *not*), determiners (*no*) or pronouns (*none*).

4.1 Clause type and status

Beyond the information from clause order and tense, **punctuation** after the sentence helps identify different types of sentences (since questions and imperatives can be an indication of the discourse-internal *speech act* relation).

For each clause, a number of **modifying adverbials** such as temporal, causal or concessive adverbials (excluding the *nachdem*- or *während*-clause), conjunctive focus adverbs (*also*, *as well*), and commentary adverbs (*doubtlessly*, *actually*, *probably*...). Additional temporal or causal adverbials, which fill the respective function for the main clause, make it less likely that the subordinate clause temporally locates or causally explains the main clause, whereas conjunctive focus adverbs often indicate a *parallel* relation. Finally commentary adverbs are indicative of discourse-internal relations since they indicate deviations from purely factual reporting.

In order to capture event contingency between clauses (which is typical for temporal and causal relations, but not for contrastive relations), we included both referential and lexico-semantical indicators: the **compatible subject pronoun** feature indicates that the subject of one clause is a compatible antecedent for the subject of the other clause (which, due to parallelism and subject preference, is a relatively robust indicator for the subjects being coreferential). In this context, morphological compatibility is relatively simple to derive from the morphological tags in the treebank (which include number and grammatical gender), but it would be expected that the same information can be reliably derived from the output of a morphological analyzer.

4.2 Shallow lexical-semantical features

In general, targeting specific linguistic properties of the clauses linked by the connective will provide crucial information in some cases (as, for example, the co-temporal reading of *während* can be excluded when tenses disagree), but is not sufficient when the choice of discourse relation is influenced by the kind of event that is denoted by the argument clauses, or more general aspects of their meaning.

Some predicates occur often enough to be used as a generalization, and often provide either linguistic hints (in the case of verbs that are typically individual-level, rather than stage-level predicates and would not be located or be used to locate temporally, e.g. *exist*) or are typically thought of as causer, or causee, of an event (as, e.g., *crash* is more likely to be the result or explanation to another event than *fly*). The **semantic head** feature includes the semantic head (i.e., main verb) of each clause, which can provide this kind of information where the main verb is informative and occurs often enough in the training data.

Since most predicates are not frequent enough to occur in a significant number, we need informative statistics that can uncover relevant aspects of their meaning. One such distributional statistic considers the type of (sub-)clauses in which verbs typically appear: verbs such as *require*, *suspect*, or *fear* often occur as part of a *because* clause, while *arrest*, *resign* or *conclude* often occur as part of a *after* adverbial clause. Bethard and Martin (2008), who use this strategy for the prediction of causal and temporal readings of *and*, are able to use n-gram search for such frequency statistics. In the case of German, morphological flexibility and the verb order in subclauses mean that it is necessary to consider a larger context. For the **association** feature in our experiments, we extracted counts from subclause occurrences in the DE-WaC corpus (Baroni and Kilgariff, 2006) using the subordinating conjunctions *bevor* (before), *nachdem* (after/as/since), *weil* (because) and *obwohl* (although). Using (local) pointwise mutual information (MI) scores, each pair of conjunction and verb lemma is assigned binary features indicating whether it has a negative score, or the quantile of lemmas for that connective, according to positive MI values.

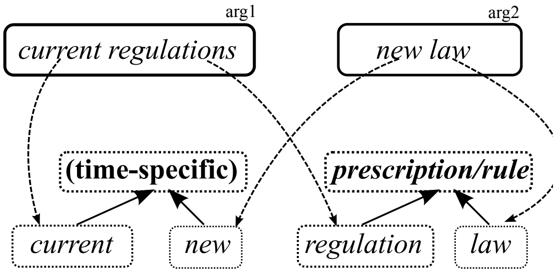

Figure 1: Lexical relation feature

The **lexical relation** feature targets pairs of words across both clauses that are taxonomically related and thus could form a contrast pair. As an example, consider *the current regulations* occurring in one clause and *the new law* in the other, which would yield a pair of time-related adjectives *current-new*, and a pair *regulation-law* of concepts that are both hyponyms of *prescription/rule* (cf. figure 1). To find these pairs of taxonomically related concepts, we use the hyperonymy hierarchy in GermaNet 5.0 (Kunze and Lemnitzer, 2002) to produce the least common subsumer of two terms plus two superordinate terms. For adjectives and verbs, requiring a least common subsumer always yields related pairs. In contrast, the upper levels of the noun hierarchy are very general, and we ensure that only related pairs are used by ignoring the upper three levels of the noun hierarchy for this feature.

Another, shallower way of representing the relation(s) between the words in each argument clause has proven to be effective in research on unlabeled relations: The **pairs of lemmas** feature extracts

All relations	dice	eq	contrast	Temp
contrast+Temporal	0.844	0.533	0.868	**0.868**
best (CRF)	0.823	0.552	0.874	0.823
best (CRF+bias)	0.855	**0.581**	0.893	0.853
best (SVMperf)	**0.857**	0.579	**0.897**	0.854
Only primary relation	Accuracy			
contrast	0.655			
baseline (CRF)	0.674			
best (CRF)	0.712			

Table 2: Results for *während*

pairs of lemmas occurring across the two argument clauses. On one hand, this feature can detect co-taxonomic pairs such as *current-new* or *rise-fall* (as well as nontaxonomic relations such as *accident-injured*) whenever these occur very frequently. On the othe hand, such a feature can also uncover the presence of a personal pronouns, or two definite articles, in each of both clauses, or particular adjectives.

Among all pairs of lemmas, we only select those that occur at least 5 times in the training data, and select the 500 most 'interesting' the by using overall entropy as a selection criterion. Using entropy in this way serves to exclude very frequent word pairs (which occur in – nearly – every pair of clauses that has been seen) as well as very infrequent ones.

4.3 Structural information

In order to account for structure, we include the **productions** feature, which is based on nonterminal and preterminal productions (e.g., `NX` → `ART ADJX NN` for an NP with a determiner, an adjective and a noun, or `ART` → `der` for *der* occurring as a determiner). Among those productions that occur in at least 500 of the clause pairs, the 500 with the highest entropy are used (filtering out those that are very rare, or frequent enough to appear in nearly *each* sentence).

5 Impact of Features

An overview on the evaluation results for *während* and *nachdem* is provided in tables 2 and 3, whereas table 4 contains more detail on the impact of each feature. In general, all of the evaluation metrics (cf. section 3.1) are improved by the rich set of features. Fine-grained accuracy (dice[2] and dice[3]) benefits more by the ranking-based CRF approach, and the best coarse-grained accuracy (eq[1] and dice[1]) is achieved by the greedy SVM classification.

Due to space reasons, we limited the feature analysis in table 4 to feature sets containing either (i) base features plus any single feature, or (ii) all but a single one of the features.

As can be seen in the table, the most difficult relations to identify are minority relations such as *contrast*, *parallel*, *evidence*, and *speech-act*. *Speech-act* is rare enough that no better-than-baseline feature set ever produces it. In contrast, the best feature set achieves F-measures of 0.41 (*contrast*), 0.39 (*parallel*) and 0.33 (*evidence*) on these relations, with precision values between 0.33 (*evidence*) and 0.36 (*contrast*), and recall values between 0.33 (*evidence*) and 0.47 (*contrast*). Considering that these relations are quite rare (the most frequent of them, *contrast*, occurs in 5.8% of the *nachdem* instances),

The feature that has most impact by itself is the presence of modifying adverbials (**mod.adv.**), especially for *parallel* and *cause* relations. The *association* feature (**assoc**) is the most effective in identifying *cause* and *evidence* relations, as it provides information on kinds of events that a verbs refers to. Co-occurrence of a verb in the sub- or main clause with the introducing or modifying connective can help to distinguish temporally-locating events (which can, e.g., occur in *before* subclauses), or states of affairs that can serve as a reason for something (which would occur in *because* or *although* subclauses).

Both of the shallow features, **productions** and lemma pairs (**wordpairs**) have a relatively broad effect and lead to successful identification of some of the minority relations (*cause*, *contrast*, *evidence*). However, they are noisy enough that overall performance drops below the baseline (in the case of word pairs, the dice measure for the finer taxonomy level and strict equality seem to improve, however).

In the reverse feature selection, however, we see that the noisy information brought in by the shallow lexical features (*productions* and *wordpairs*) is quite useful: performance drops very visibly without these features (0.844 to 0.835 for removing the *productions* feature, to 0.817 for *wordpairs*).

Looking at the learning curves (for the full feature set minus the *assoc* feature), in figure 2, we find that the identification of *cause* and *enable* relations seems to be relatively robust to sparse data problem, as the improvement from 20% of training data (i.e., randomly subsampling each train-

setting	dice[1]	dice[2]	dice[3]	eq[1]	Comparison	Result	Temporal	contrast	cause	evidence
random	0.742	0.707	0.627	0.415	0.065	0.610	0.938	0.000	0.231	0.083
Temporal+enable	0.829	0.789	0.680	0.541	0.000	0.752	0.968	0.000	0.000	0.000
baseline (CRF)	0.782	0.747	0.683	0.466	0.143	0.625	0.953	0.087	0.248	0.211
best (CRF)	0.823	0.806	**0.729**	0.548	0.341	0.678	**0.974**	0.333	0.355	**0.400**
best (CRF+bias)	0.845	**0.814**	0.710	0.595	0.348	0.764	0.972	0.286	0.347	0.286
baseline (SVMperf)	0.829	0.789	0.680	0.541	0.000	0.752	0.968	0.000	0.000	0.000
best (SVMperf)	**0.849**	0.811	0.718	**0.609**	**0.514**	**0.763**	0.970	**0.410**	**0.369**	0.333

Table 3: Results for *nachdem*

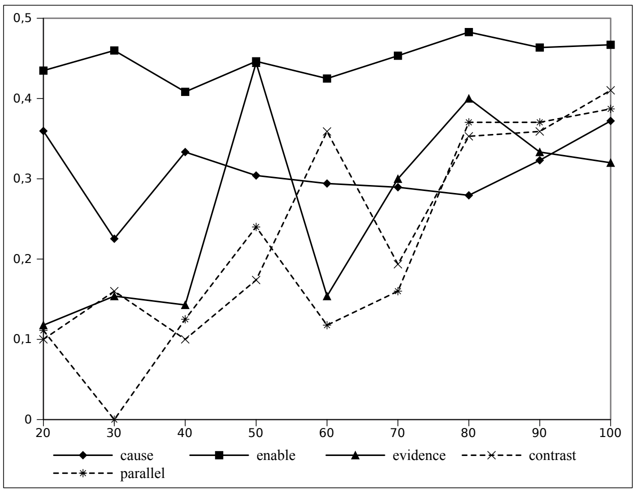

Figure 2: Learning curves for single relations (*nachdem* only)

ing fold to 20% of its size) to the complete data only yields limited improvement, whereas relations such as *evidence*, *contrast* and *parallel* seem to profit strongly from more data (which is understandable, however, since these relations are less frequent than the others).

Although the annotated instances stem from a relatively large corpus (slightly over one million words), it seems very plausible that larger training data would benefit the disambiguation results. For connective annotation on a fixed-size corpus (such as the TüBa-D/Z, or the Penn Treebank used for the Penn Discourse Treebank), combining the benefits of connective-specific and non-specific disambiguation would be especially relevant, as the former allows to model the specific connective meaning, whereas connective-independent models would be less sensitive to sparse data.

6 Summary

We carried out multilabel tagging experiments on two datasets: one containing occurrences of *nachdem* from the TüBa-D/Z corpus (shown in table 3), one containing occurrences of *während*, using 10-fold cross-validation on the training set. For both the CRF-based approach and the SVM-based one-versus-all reduction, the best-performing feature set we found contains all features minus the *association* feature.

For both *nachdem* and *während*, the most frequent sense (Temporal+enable, or Temporal+contrast) is by far predominant and yields a very strong baseline, which the CRF-based classifier only surpasses for *nachdem* with an appropriate setting for the bias parameter to prevent the classifier from under-labeling (i.e., assigning fewer relations than optimal). Both the biased CRF classifier and the greedy SVM-based approach outperform the most-frequent sense baseline for all aggregate measures, which is more difficult for the top level of the taxonomy where one single coarse-grained relation combination often accounts for over 50% of all instances.

To our knowledge, this study is the first successful study on disambiguating German connectives, after the results of (Bayerl, 2004) who studied the explicit connective *wenn* (if/when), which stay further below the most-frequent sense baseline. We take this to confirm the intuition that problems in large-scale discourse classification, including those thought to be unrewarding such as ambiguous explicit connectives, are best tackled with a combination of an annotation scheme that is appropriate to the task (i.e., focused on coherence relations rather than speaker intentions), informative features, and a machine learning approach that can make use of these features to reproduce all the distinctions that are present in the annotation.

We also hope that the general direction of (i) reproducing all of the information present in the gold annotation and (ii) using a rich set of features for the disambiguation of ambiguous explicit con-

	dice[1]	dice[2]	dice[3]	equal	Comp.	contr.	parallel	Result	cause	enable	evidence	sp.-act	Temp.
base (cl. order, tense, neg.)	0.829	0.789	0.678	0.541	0.000	0.000	0.000	0.752	0.054	0.485	0.000	0.000	0.968
base + assoc	0.809	0.768	0.676	0.507	0.075	0.000	0.067	0.728	**0.338**	0.477	**0.276**	0.000	0.968
base + csubj	0.829	0.789	0.678	0.541	0.000	0.000	0.000	0.751	0.073	**0.488**	0.000	0.000	0.968
base + sem.head	0.829	0.789	0.680	0.541	0.000	0.000	0.000	0.752	0.103	0.485	0.000	0.000	0.968
base + lexrel	0.829	0.789	0.678	0.541	0.000	0.000	0.000	0.752	0.133	0.485	0.000	0.000	0.968
base + mod.adv.	**0.832**	0.789	0.675	0.551	0.216	0.000	**0.222**	**0.753**	0.162	0.484	0.000	0.000	0.968
base + productions	0.782	0.731	0.645	0.480	**0.272**	0.159	0.150	0.700	0.331	0.429	0.105	**0.111**	0.949
base + punc	0.827	0.789	0.680	0.541	0.000	0.000	0.000	0.749	0.056	**0.488**	0.000	0.000	0.968
base + wordpairs	0.824	0.774	**0.683**	**0.551**	0.262	**0.294**	0.000	0.741	0.284	0.458	0.261	0.000	0.965
all	0.844	0.802	0.706	0.588	0.478	0.343	0.312	0.756	0.356	0.430	0.435	0.000	0.970
all w/o assoc	**0.849**	**0.811**	**0.718**	**0.609**	**0.514**	**0.410**	0.387	**0.763**	0.369	**0.463**	0.333	0.000	0.970
all w/o csubj	0.835	0.795	0.703	0.568	0.485	0.343	0.323	0.736	0.333	0.431	**0.455**	0.000	0.970
all w/o sem.head	0.844	0.802	0.701	0.588	0.478	0.333	0.323	0.758	0.338	0.423	0.381	0.000	0.968
all w/o lexrel	0.843	0.799	0.710	0.588	0.507	0.343	**0.389**	0.753	**0.385**	0.442	0.364	0.000	0.968
all w/o mod.adv.	0.840	0.803	0.699	0.585	0.386	0.375	0.160	0.754	0.333	0.419	0.435	0.000	0.968
all w/o productions	0.834	0.781	0.689	0.575	0.486	0.350	0.353	0.738	0.281	0.400	0.400	0.000	0.964
all w/o punc	0.842	0.800	0.706	0.585	0.478	0.343	0.312	0.754	0.365	0.432	0.435	0.000	0.968
all w/o wordpairs	0.817	0.769	0.676	0.541	0.465	0.302	0.242	0.721	0.362	0.408	0.244	0.000	0.953

Table 4: Impact of features (for *nachdem*, SVMperf)

nectives will be a fruitful direction for discourse relation disambiguation also in other languages than German.

References

Baroni, M. and Kilgariff, A. (2006). Large linguistically-processed web corpora for multiple languages. In *EACL 2006*.

Bayerl, P. S. (2004). Disambiguierung deutschsprachiger Diskursmarker: Eine Pilot-Studie. *Linguistik Online*, 18.

Bethard, S. and Martin, J. (2008). Learning semantic links from a corpus of parallel temporal and causal relations. In *ACL/HLT 2008*.

Joachims, T. (2005). A support vector method for multivariate performance measures. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Kunze, C. and Lemnitzer, L. (2002). GermaNet – representation, visualization, application. In *Proceedings of LREC 2002*.

Lin, Z., Kan, M.-Y., and Ng, H. T. (2009). Recognizing implicit discourse relations in the Penn Discourse Treebank. In *EMNLP 2009*.

Marcu, D. and Echihabi, A. (2002). An unsupervised approach to recognizing discourse relations. In *ACL 2002*.

Miltsakaki, E., Dinesh, N., Prasad, R., Joshi, A., and Webber, B. (2005). Experiments on sense annotations and sense disambiguation of discourse connectives. In *TLT 2005*.

Miyao, Y. and Tsujii, J. (2002). Maximum entropy estimation for feature forests. In *HLT 2002*.

Pitler, E., Louis, A., and Nenkova, A. (2009). Automatic sense prediction for implicit discourse relations in text. In *ACL-IJCNLP 2009*.

Pitler, E. and Nenkova, A. (2009). Using syntax to disambiguate explicit discourse connectives in text. In *ACL 2009 short papers*.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse Treebank 2.0. In *Proceedings of LREC 2008*.

Soricut, R. and Marcu, D. (2003). Sentence level discourse parsing using syntactic and lexical information. In *Proc. HLT/NAACL-2003*.

Sporleder, C. and Lascarides, A. (2008). Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14(3):369–416.

Telljohann, H., Hinrichs, E. W., Kübler, S., Zinsmeister, H., and Beck, K. (2009). Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technical report, Seminar für Sprachwissenschaft, Universität Tübingen.

Versley, Y. and Rehbein, I. (2009). Scalable discriminative parsing for German. In *Proc. IWPT 2009*.

Wellner, B. (2009). *Sequence Models and Ranking Methods for Discourse Parsing*. PhD thesis, Brandeis University.